Model-based Clustering: Clustering with Probabilistic Mixture Models

Ahmed BADI
ahmedbadi905@gmail.com
linkedin.com/in/badi-ahmed

January 3, 2026

Abstract

Model-based clustering approaches the clustering problem from a probabilistic perspective: instead of directly grouping points by distance, it assumes that the data were generated from a mixture of underlying probability distributions, and it infers clusters by estimating the parameters of this mixture [1], [2]. The most widely used instance is the Gaussian Mixture Model (GMM), in which each cluster corresponds to one multivariate normal component characterized by a mean vector and covariance matrix. This article introduces the core ideas of model-based clustering, presents the Gaussian Mixture Model and the Expectation–Maximization (EM) algorithm used to fit it, and explains how the resulting soft cluster assignments differ from hard methods like k-means. We also discuss model selection for choosing the number of components, extensions such as Bayesian mixtures and Dirichlet process mixtures, and the strengths and limitations of model-based approaches in practice.

Keywords: Model-based Clustering, Gaussian Mixture Model, EM Algorithm, Soft Clustering, BIC, Mixture Models.

1 Introduction

In distance-based clustering methods such as k-means, clusters are defined implicitly through geometric criteria: points are assigned to the nearest centroid, and the algorithm optimizes a variance-based objective [3]. By contrast, **model-based clustering** starts from an explicit probabilistic model of the data-generating process.

The central assumption is that the observed data $X = \{x_1, \ldots, x_n\}$ are independent draws from a mixture of K component distributions:

$$p(x) = \sum_{k=1}^{K} \pi_k \, f(x \mid \theta_k), \tag{1}$$

where:

- π_k are non-negative *mixing proportions* with $\sum_{k=1}^{K} \pi_k = 1$,
- $f(x \mid \theta_k)$ is the probability density (or mass) function of component k with parameters θ_k,
- each component is interpreted as one cluster.

Clustering then consists of two linked tasks:

1. Estimate the mixture parameters $\{\pi_k, \theta_k\}_{k=1}^{K}$ from the data.
2. Assign each data point to a component—either *softly* via posterior probabilities, or *hardly* by choosing the most probable component.

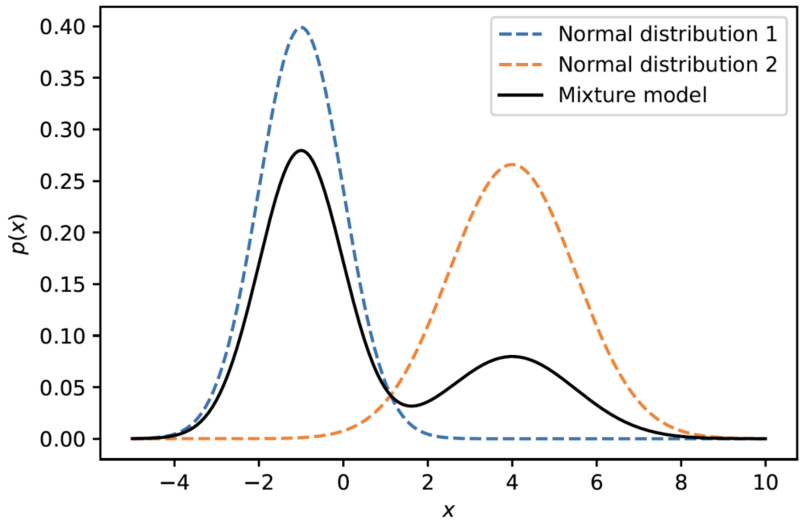

Figure 1: One-dimensional Gaussian mixture: the overall density (black) is the sum of several component Gaussians (colored), each corresponding to a cluster.

The most common choice for f in continuous spaces is the multivariate Gaussian distribution, yielding a **Gaussian Mixture Model** (GMM).Other distributions can be used for discrete data (e.g., multinomial or Poisson mixtures).

2 Gaussian Mixture Models (GMM)

2.1 Definition

In a d-dimensional feature space, a GMM with K components assumes:

$$p(x) = \sum_{k=1}^{K} \pi_k \mathcal{N}(x \mid \mu_k, \Sigma_k), \tag{2}$$

where $\mathcal{N}(x \mid \mu_k, \Sigma_k)$ is the multivariate normal density with mean $\mu_k \in \mathbb{R}^d$ and covariance matrix $\Sigma_k \in \mathbb{R}^{d \times d}$.

Each component (cluster) is thus characterized by:

- **Mean** μ_k: center of the cluster.
- **Covariance** Σ_k: shape, orientation, and spread.

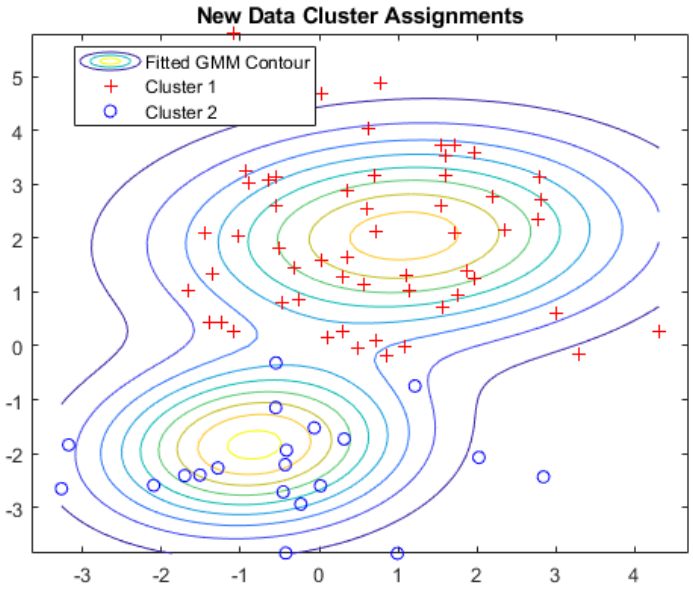

Figure 2: Gaussian mixture in 2D: each component defines an elliptical contour; overlapping components result in soft cluster boundaries.

Because covariance matrices allow ellipsoidal shapes, GMMs can model clusters with different sizes, orientations, and degrees of overlap, which is not possible with simple spherical assumptions.

2.2 Latent Variables and Soft Assignments

Model-based clustering introduces latent variables Z_i indicating which component generated each observation x_i. For K components, Z_i is often encoded as a one-hot vector $(Z_{i1}, \ldots, Z_{iK})$ with:

$$Z_{ik} = \begin{cases} 1 & \text{if } x_i \text{ belongs to component } k, \\ 0 & \text{otherwise.} \end{cases}$$

Under the mixture model, the posterior probability that x_i comes from component k is:

$$\gamma_{ik} = P(Z_{ik} = 1 \mid x_i) = \frac{\pi_k \mathcal{N}(x_i \mid \mu_k, \Sigma_k)}{\sum_{j=1}^{K} \pi_j \mathcal{N}(x_i \mid \mu_j, \Sigma_j)}. \tag{3}$$

The vector $(\gamma_{i1}, \ldots, \gamma_{iK})$ provides a **soft clustering**: x_i has a graded membership in each cluster, instead of a single hard label.

3 Parameter Estimation with EM

The parameters $\{\pi_k, \mu_k, \Sigma_k\}$ are typically estimated by **maximum likelihood**. However, maximizing the incomplete-data log-likelihood directly is difficult because of the latent variables Z_i. The standard solution is the **Expectation–Maximization (EM)** algorithm.

3.1 EM Algorithm for GMM

EM alternates between:

E-step: Compute responsibilities γ_{ik} using current parameters (Eq. (3)).

M-step: Update parameters by maximizing the expected complete-data log-likelihood given γ_{ik}.

The M-step updates are:

$$N_k = \sum_{i=1}^{n} \gamma_{ik}, \tag{4}$$

$$\pi_k^{\text{new}} = \frac{N_k}{n}, \tag{5}$$

$$\mu_k^{\text{new}} = \frac{1}{N_k} \sum_{i=1}^{n} \gamma_{ik} x_i, \tag{6}$$

$$\Sigma_k^{\text{new}} = \frac{1}{N_k} \sum_{i=1}^{n} \gamma_{ik} (x_i - \mu_k^{\text{new}})(x_i - \mu_k^{\text{new}})^\top. \tag{7}$$

Algorithm 1 EM for Gaussian Mixture Models

Require: Data X, number of components K

1: Initialize $\{\pi_k, \mu_k, \Sigma_k\}_{k=1}^{K}$ (e.g., k-means or random).
2: **repeat**
3: **E-step:** For each i, k, compute responsibilities γ_{ik} (Eq. (3)).
4: **M-step:** Update π_k, μ_k, Σ_k using the equations above.
5: **until** log-likelihood converges or max iterations reached

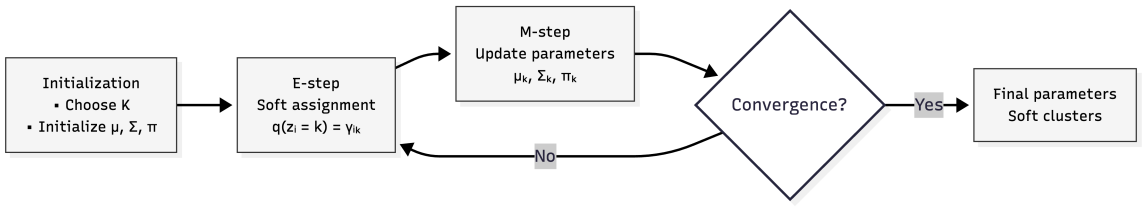

Figure 3: EM for GMM: initialization, soft assignment of points to components (E-step), and parameter updates (M-step) over iterations.

EM is guaranteed to increase (or leave unchanged) the log-likelihood at each iteration and converges to a local maximum of the likelihood function.

4 Comparing GMM and k-means

GMMs and k-means are closely related:

- If all covariances are constrained to be spherical and equal ($\Sigma_k = \sigma^2 I$) and mixing proportions are equal, EM for GMM reduces to a soft version of k-means.
- Hard cluster assignments can be obtained from GMM by taking $\arg\max_k \gamma_{ik}$, analogous to k-means labels.

However, GMMs offer several important advantages:

- **Soft clustering:** Each point has membership probabilities, useful when clusters overlap.
- **Flexible shapes:** Full covariance matrices allow elliptical clusters with different orientations and scales.

- **Probabilistic interpretation:** The model provides likelihoods, which can be used for density estimation and anomaly detection.

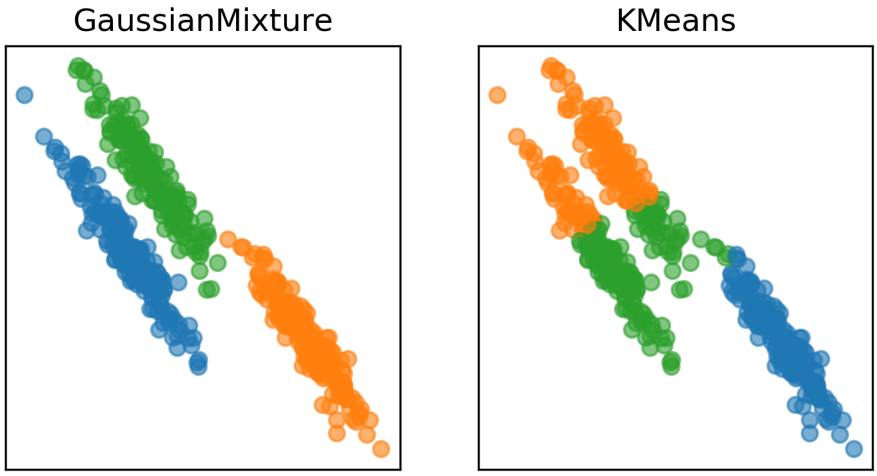

Figure 4: Clustering elongated Gaussian clusters: k-means (left) struggles with non-spherical shapes, while GMM with full covariances (right) adapts ellipses to each cluster.

Limitations include:

- Sensitivity to initialization and local optima (like k-means).
- Potential overfitting, especially with full covariance matrices in high dimensions.
- Computational cost for large K or high d.

5 Choosing the Number of Components

A crucial question in model-based clustering is selecting K. Common strategies rely on penalized likelihood criteria:

- **BIC** (Bayesian Information Criterion):

$$\text{BIC} = -2 \log L_{\max} + p \log n,$$

  where $L_{\max}$ is the maximized likelihood, p is the number of free parameters, and n is the number of observations.

- **AIC** (Akaike Information Criterion), similar but with a different penalty.

BIC is widely used in GMM clustering to jointly estimate parameters and choose K.

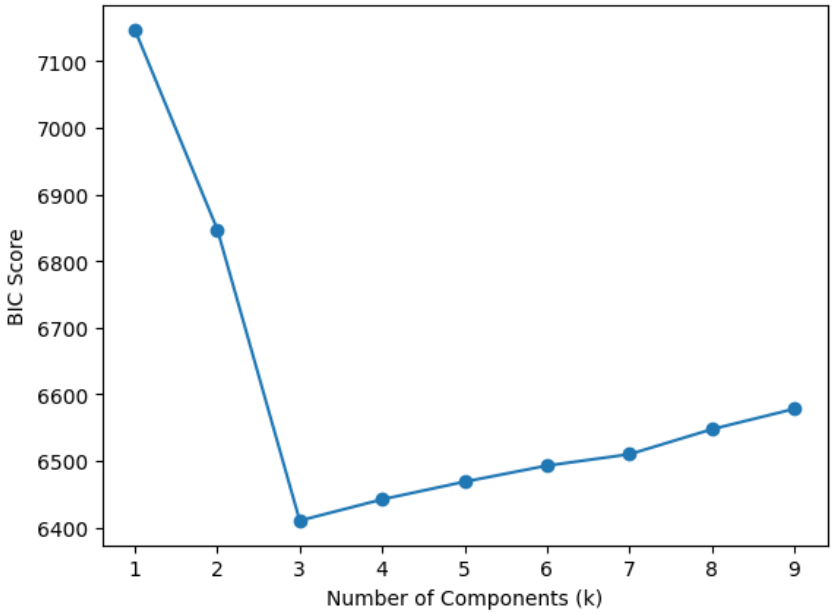

Figure 5: Model selection with BIC: the best number of components corresponds to the minimum BIC value.

There is an important conceptual distinction: the number of *mixture components* selected for the best density model may not equal the most intuitive number of *clusters*.In practice, components can sometimes be merged or interpreted jointly to form higher-level clusters.

6 Beyond Finite GMMs

6.1 Other Component Distributions

While Gaussian mixtures dominate continuous data applications, model-based clustering extends naturally to other distributions [2]:

- Multinomial mixtures for text and categorical data (latent class models).
- Poisson or negative binomial mixtures for count data.
- Mixtures of specialized distributions in domains like genomics or finance.

6.2 Bayesian Mixtures and Dirichlet Process Models

Bayesian approaches place priors on mixture parameters and sometimes on K itself.A Dirichlet process mixture model (DPMM) can be seen as a mixture with an unbounded number of components, where the effective number of occupied components is inferred from the data rather than fixed in advance.

These models:

- Provide full posterior uncertainty over cluster assignments.
- Can automatically adapt the number of clusters.
- Require more advanced inference methods (e.g. MCMC, variational inference).

7 When to Use Model-based Clustering?

Model-based clustering is particularly useful when:

- Clusters are approximately well-described by parametric distributions (e.g. Gaussian).
- Overlapping clusters are expected and soft assignments are valuable.
- A probabilistic interpretation (densities, likelihoods) is needed.
- One wants to use the same model both for clustering and for density-based tasks such as anomaly detection or simulation.

It may be less appropriate when:

- Cluster shapes are highly non-Gaussian and cannot be captured by a small number of components.
- Dimensionality is extremely high and sample size is limited, leading to unstable covariance estimates.
- Computational resources are constrained and simpler methods (e.g. k-means) are sufficient.

8 Conclusion

Model-based clustering frames clusters as components of a generative probabilistic model. Gaussian Mixture Models combined with the EM algorithm provide a flexible and interpretable way to perform soft clustering, allowing clusters to overlap and take on arbitrary ellipsoidal shapes.Penalized likelihood criteria such as BIC support principled selection of the number of components, while Bayesian extensions further quantify uncertainty and can infer K automatically.

When their assumptions are approximately satisfied, model-based methods can outperform purely geometric clustering, especially in complex, overlapping data where probabilities and uncertainty matter. Understanding both their power and their limitations helps decide when they are the right tool in the broader clustering toolbox.

References

[1] C. Fraley and A. E. Raftery, "Model-based clustering, discriminant analysis, and density estimation," *Journal of the American Statistical Association*, vol. 97, no. 458, pp. 611–631, 2002.

[2] G. McLachlan and D. Peel, *Finite Mixture Models*. Wiley, 2000.

[3] T. Hastie, R. Tibshirani, and J. Friedman, *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. Springer, 2009.